**Lab 3** **Statistics for Laboratory Scientists**
**Due: 15 March 2006**

## Introduction

In this lab, we will analyze two small datasets, in order to illustrate the use of R for calculating confidence intervals and performing t-tests. One data set will be used for illustration. The analysis of the second data set will largely be left up to you, although code will given to assist you with the initial bits.

The code for this lab will again be available at the following webpage.

`http://www.biostat.jhsph.edu/~kbroman/teaching/labstat/third/labs.html`

The file `lab3.R` contains all of the code. As with the first two labs, you may wish to download this file and open it within R using, from the menu bar, File → Display file, so that you may copy and paste commands from the file into the R console window.

## Hummer et al.

The first data set is taken from BT Hummer, X-L Li, BA Hassel (2001) Role for p53 in gene induction by double-stranded RNA. *J. Virol.* **75**: 7774–7777. The data are contained in the file `hummer.csv`. Load the file directly into R with the following command.

```
hummer <- read.csv("http://www.biostat.jhsph.edu/~kbroman/teaching/data/hummer.csv")
```

Type `ls()` to see that the data is now in your workspace. Type `hummer` to view the data. These data concern promoter activity following treatment with IFN or SV in the presence or absence of p53. The first column of these data give the treatment. The second column indicates whether p53 was present or absent. The last column contains the promoter activity measurements. Note: I must admit that I don't completely understand the experiments used to generate either of the data sets that we are considering in this lab. If you are interested, please see the cited papers.

### *Plot the data*

The first thing you should do after obtaining new data (and getting it into R): **plot them** (or it)! My favorite way to plot the sort of two-sample data that we are studying here is the "dotplots" that you've seen numerous times in lecture. The built-in function in R for making this sort of plot is `stripchart()`, but I don't like this function very much, so I've written an alternate version, `dotplot()`. This function appears on appears on the second-to-last page in the lab, and at the beginning of the file `lab3.R` (that you should have downloaded and opened within R). Copy and paste the text for this function into R, so that you may use it on the data.

The authors were most interested in comparing the promoter activity after a treatment in the presence vs the absence of p53. Let us first look at the IFN treatment. We want to pull out the promoter activity measurements in the absence of p53 as `x` and in the presence of p53 as `y`. There are two ways to do this. First, the direct way:

```
x <- hummer[4:6,3]
y <- hummer[1:3,3]
```

Alternatively, we can do the following:

```
x <- hummer[hummer[,1]=="IFN" & hummer[,2]=="--", 3]
y <- hummer[hummer[,1]=="IFN" & hummer[,2]=="++", 3]
```

Now we can plot the data using `dotplot()`. This function will automatically place 95% confidence intervals for the population means next to each sample. We'll see how to calculate those shortly.

```
dotplot(x,y)
```

We can do the same thing for the SV treatment.

```
x2 <- hummer[hummer[,1]=="SV" & hummer[,2]=="--", 3]
y2 <- hummer[hummer[,1]=="SV" & hummer[,2]=="++", 3]
dotplot(x2,y2)
```

***Calculating a confidence interval for a population mean***

To calculate a 95% confidence interval for a population mean on the basis of data, one must calculate (a) the sample mean, (b) the sample SD, and (c) the appropriate quantile from the t distribution with the appropriate degrees of freedom. In R, you can use the function `t.test()` to do this painlessly.

For example, to get a 95% confidence interval for the "true" mean promoter response after treatment with IFN and in the absence of p53, type the following. (Note that `x` was created from the appropriate portion of the `hummer` data, as shown above.)

```
t.test(x)
```

The output of this function includes the relevant confidence interval. It also contains a bunch of other junk that you probably don't care about.

We can get confidence intervals for the mean promoter activity after the treatment with IFN in the presence of p53, and after treatment with SV in the absence and presence of p53, as follows.

```
t.test(y)
t.test(x2)
t.test(y2)
```

These confidence intervals are what are plotted when you use `dotplot`. You can learn more about the function `t.test` in its help file. Type the following.

```
?t.test
```

Note, in particular, that you can obtain confidence intervals with different "levels of confidence" using the argument `conf.level`. For example, to get a 99% confidence for the mean promoter activity in the absence of p53, type the following.

```
t.test(x, conf.level=0.99)
```

***Confidence interval for the difference between population means***

The same function, `t.test()`, may be used to compare two samples. If we wish to calculate a 95% confidence interval for the difference between the mean promoter activity after IFN treatment, in the presence and absence of p53, we could type the following. (Again, `x` and `y` were created above.)

```
t.test(x, y)
```

In the output, you'll see the 95% confidence interval for the difference.

Notice what happens when you switch the order of `x` and `y`:

```
t.test(y, x)
```

Note that this confidence interval is obtained allowing that the two populations have possibly different standard deviations (SDs). You can use the argument `var.equal` to get the confidence interval that assumes a common SD for the two populations, and uses the pooled estimate of that SD.

```
t.test(x, y, var.equal=TRUE)
```

Also calculate a 95% confidence interval for the difference in the average promoter activity after SV treatment, in the presence and absence of p53.

```
t.test(x2, y2)
```

***Testing for a difference between population means***

I'm sure that you noticed that the `t.test()` function not only gives you a confidence interval for the difference between population means, but also gives you a P-value for the test of whether the difference between the means is strictly 0.

By default, `t.test()` performs a two-tailed test. You can get a one-tailed test using the argument `alternative`, which should be set to either `"two.sided"`, `"less"`, or `"greater"`. Try the following:

```
t.test(x2, y2, alternative="two.sided")
t.test(x2, y2, alternative="less")
t.test(x2, y2, alternative="greater")
```

The argument `alternative` refers to whether the mean of the population from which the first sample was drawn is different, less than, or greater than that of the second sample.

Note that the P-value for testing the difference between the mean promoter activity after IFN treatment, in the presence and absence of p53, is 0.44. Thus we conclude that the observed difference could reasonably be due to chance variation.

On the other hand, the P-value for testing the difference between the mean promoter activity after SV treatment, in the presence and absence of p53, is 0.04. Thus we conclude that the observed difference *cannot* reasonably be explained by chance variation: the difference is "statistically significant."

#### *Another way to use* `t.test()`

There is another way to use the function `t.test()` for these data that is in some ways simpler and in other ways more complex.

If we wish to compare the difference in the mean promoter activity after IFN treatment, in the presence and absence of p53, we could type the following.

```
t.test(activity ~ p53, data=hummer, subset=(medium=="IFN"))
```

This means: do a t-test to compare the values in the column `activity` when split into two groups according to the values in the column `p53`. The argument `data=hummer` tells `t.test()` to consider the data set `hummer`. The argument `subset=(medium=="IFN")` tells it to consider only those *rows* for which the value in the *column* `medium` are equal to `"IFN"`.

We can do the same thing for the SV treatment.

```
t.test(activity ~ p53, data=hummer, subset=(medium=="SV"))
```

## Berrios et al.

The second data set is taken from JC Berrios, MA Schroeder, RD Hubmayr (2001) Mechanical properties of alveolar epithelial cells in culture. *J. Appl. Physiol.* **91**: 65–73. The data concern cell surface adhesion receptor expression after 1 and 4 days in culture.

The data are contained in the file `berrios.csv`. Load the file directly into R with the following command.

```
berrios <- read.csv("http://www.biostat.jhsph.edu/~kbroman/teaching/data/berrios.csv")
```

If you type `berrios`, you should see the data. There are 5 measurements for day 1 and 7 for day 4. The first column is the day; the second column contains the expression intensity measurements.

Create objects `x` and `y` containing the data for days 1 and 4, respectively, as follows:

```
x <- berrios[berrios[,1]==1, 2]
y <- berrios[berrios[,1]==4, 2]
```

Plot the data:

```
dotplot(x,y)
```

Use the function `t.test()` to answer the questions on the last page of the lab.

### Code for the `dotplot()` function

```
dotplot <-
function(x,y,includeCI=TRUE)
{
  # Arrange the data
  X <- c(x,y)
  Y <- rep(1:0,c(length(x),length(y)))

  # jitter the Y positions
  Y <- Y + runif(length(Y),-0.1,0.1)

  # If requested, calculate the CI's
  # and make sure x-limits allow plot of CI's
  if(includeCI) {
    xci <- t.test(x)$conf.int
    yci <- t.test(y)$conf.int
    xlimits <- range(c(x,y,xci,yci))
  }
  else xlimits <- range(c(x,y))

  # make plot
  plot(X,Y,ylim=c(-0.5,1.5),yaxt="n",lwd=2,xlab="",ylab="",
       xlim=xlimits)
  abline(h=0:1,lty=2,col="gray")

  # add Y-axis labels
  u <- par("usr")
  segments(u[1],0:1,u[1]-diff(u[1:2])*0.03,0:1,xpd=TRUE)
  text(u[1]-diff(u[1:2])*0.08,1:0,c("A","B"),xpd=TRUE,cex=1.3)

  # add confidence intervals, if requested
  if(includeCI) {
    segments(xci[1],1.2,xci[2],1.2,lwd=2,col="blue")
    segments(xci,1.18,xci,1.22,lwd=2,col="blue")
    segments(mean(x),1.15,mean(x),1.25,lwd=2,col="blue")

    segments(yci[1],0.2,yci[2],0.2,lwd=2,col="red")
    segments(yci,0.18,yci,0.22,lwd=2,col="red")
    segments(mean(y),0.15,mean(y),0.25,lwd=2,col="red")
  }
}
```

**Questions to be answered**

*Concerning the second data set, from Berrios et al. (2001):*

1. Calculate a 95% confidence interval for the average intensity at day 1.
2. Calculate a 95% confidence interval for the average intensity at day 4.
3. Calculate a 95% confidence interval for the difference between the average intensities at days 1 and 4.
4. Calculate the P-value for the test of the difference between the average intensities at days 1 and 4.
5. What do you conclude from these data?